DOCUMENT RESUME

ED 447 211 TM 032 119

AUTHOR Henson, Robin K.
TITLE Sacrificing Reliability and Exalting Sampling Error at the Altar of Parsimony: Some Cautions Concerning Short-Form Test Development.
PUB DATE 2000-11-16
NOTE 29p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (Bowling Green, KY, November 15-17, 2000).
PUB TYPE Opinion Papers (120) -- Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Factor Structure; Psychometrics; *Reliability; *Sampling; *Test Construction; Test Format; Test Length
IDENTIFIERS *Parsimony (Statistics)

ABSTRACT

The purpose of this paper is to highlight some psychometric cautions that should be observed when seeking to develop short form versions of tests. Several points are made: (1) score reliability is impacted directly by the characteristics of the sample and testing conditions; (2) sampling error has a direct influence on reliability and factor structure of scores; and (3) caution should be used when developing short forms of tests when non-normative samples are used. (Contains 3 tables and 32 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

Running head: SACRIFICING RELIABILITY

ED 447 211

Sacrificing Reliability and Exalting Sampling Error at the Altar of Parsimony: Some Cautions Concerning Short-Form Test Development

Robin K. Henson

University of North Texas 76203-1337

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Robin Henson
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
1

TM032119

Paper presented at the annual meeting of the Mid-South Educational Research Association, November 16, 2001, Bowling Green, KY. Correspondence concerning this manuscript should be sent to the author at rhenson@tac.coe.unt.edu.

Abstract

The purpose of the present paper is to highlight some psychometric cautions when seeking to develop short form versions of tests. Several points are made: a) score reliability is directly impacted by the characteristics of the sample and testing conditions, b) sampling error directly influences reliability and factor structure of scores, and c) caution should be used when developing short forms of tests when using non-normative samples.

Sacrificing Reliability and Exalting Sampling Error at the Altar of Parsimony: Some Cautions Concerning Short-Form Test Development

The principle of parsimony is a time honored tradition in science. The principle holds that given two seemingly equally valid explanations for a phenomenon, the most straightforward or simple explanation is most likely true. In research, the goal of parsimony often manifests itself in researchers' attempts to explain the most (perhaps the most variance between two sets of variables via canonical correlation) with the fewest number of variables (perhaps the fewest predictors). In an important summative article concerning his experiences in research practice, Jacob Cohen (1990) advocated for parsimony and claimed that "less is more, except of course for sample size" and "simple is better" (p. 1304, 1305; emphasis in original).

This is both a noble and practical goal. Parsimony can virtuously serve as a check against research on multiple variables that has no theoretical rationale to explain the variables' relationships. Practically, researchers also often hope to parsimoniously collect the most information at the least cost and effort, sometimes due to limited resources or access to subjects. Social psychology has the mini-max principle as its own generalization of parsimony (Myers, 1990). The mini-max

principle holds that, in social contexts, persons seek to minimize costs while maximizing benefits. In research, there are often practical matters of data collection that force this perspective.

More specifically, test development is directly impacted by the need to gain the most information for the least cost/effort. For example, test developers often seek to develop the shortest possible test that will still yield reliable and valid scores. Test-taker fatigue, time, and cost of test publication and administration are thusly minimized. Accordingly, many tests (both commercially published instruments and non-commercial data collection scales) are reduced to short form versions for use with future persons and samples.

The purpose of the present paper is to highlight some psychometric cautions when seeking to develop short form versions of tests. Several points are made: a) score reliability is directly impacted by the characteristics of the sample and testing conditions, b) sampling error directly influences reliability and factor structure of scores, and c) caution should be used when developing short forms of tests when using non-normative samples.

What Impacts Reliability Estimates?

It has been repeatedly argued that reliability is a function of scores and not tests (see e.g., Pedhazur &

Schmelkin, 1991; Thompson, 1994; Vacha-Haase, 1998). Furthermore, the scores obtained on a given test are also dependent on the characteristics of the sample tested, and perhaps, the testing conditions and other measurement features. As Thompson (1994) correctly noted: "The same measure, when administered to more heterogeneous or more homogeneous sets of subjects, will yield scores with differing reliability" (p. 839).

For example, if we assume that a sample is heterogeneous as regards the trait of interest, then the subjects will likely score differently from each other (at least to the degree of heterogeneity assumed). However, if we assume that these persons are homogeneous on the trait of interest, then they will score similarly. A hypothetical case will be used to illustrate this dynamic. Table 1 presents scores for six persons on five items when assuming these two conditions. Items are scored right (1) or wrong (0). As expected, these data yielded a larger total score variance and coefficient alpha for the heterogeneous group ($\sigma^2 = 3.50$; $\alpha = .83$) than the homogeneous group ($\sigma^2 = .30$; $\alpha = -5.00$).

INSERT TABLE 1 ABOUT HERE

The reliability estimates in Table 1 are a function of the ratio between the sum of the item variances and the total test variance. This ratio is found in the coefficient alpha formula:

$$\alpha = \underline{k}/(\underline{k}-1)[1-(\Sigma\sigma_{ITEM}^2/\sigma_{TOTAL}^2)],$$

where $\underline{k}$ is the number of items on the test, σ_{ITEM}^2 is the individual variance for each item, and σ_{TOTAL}^2 is the variance of the composite total test scores (Anastasi & Urbina, 1997).

Generally, as the sum of the item variances decreases and the total test score variance increases, internal consistency reliability estimates will increase. Cronbach's coefficient alpha (Cronbach, 1951) is based on this ratio, and since alpha is a generalization of KR-20 (Kuder & Richardson, 1937), it follows that KR-20 also uses the item to total variance ratio. Readers are referred to Henson (2000), Reinhardt (1996), and Thompson (1999) for accessible treatments of this ratio including explanations of why alpha can have a negative value as we found above. (This is paradoxical because alpha is a variance-accounted-for statistic in a squared metric.)

Because homogeneous samples will yield lower total variance, tests given to such samples will tend to yield lower reliability estimates. This clearly is a function of the

characteristics of the sample and not the test per se. As such, Reinhardt (1996) explained that "both the characteristics of the person sample selected and the characteristics of the test item can affect coefficient alpha" (p. 6). Furthermore, Dawis (1987) emphasized that "reliability is a function of sample as well as of instrument, [reliability] should be evaluated on a sample from the intended target population - an obvious but sometimes overlooked point" (p. 486). The point is clear - score reliability may vary depending on the characteristics of the sample from which the scores are obtained. It logically follows, then, that reliability should be estimated each time a test is administered, because sample compositions may vary across test administrations.

It is commonly assumed that reliability estimates can be increased by adding items to a test. This may or may not be true, because, as noted above, the central element determining the magnitude of the reliability estimate is the ratio between item to total variance. If adding items to a test positively impacts this ratio (i.e., the item increases total test variance more than it increases the sum of the item variances), then alpha will indeed increase. In addition, if the items added are of at least equal quality with the items on the original test but do not change the ratio, then the items will still bolster overall alpha because the formula for alpha contains a

correction term [$\underline{k}/(\underline{k}-1)$] for item sampling bias. However, the correctional power of this term will decrease as the number of items increases (Reinhardt, 1996).

Sacrificing Reliability

When developing short form versions of tests (often for the sake of parsimony for whatever reason), one risk to reliability is quite simply the reduced number of items on the test. In general, as a test gets shorter, it will yield less reliable scores. It should be emphasized again, however, that the most central element for reliability estimates is the item to total variance ratio. Therefore, not all short forms of a test will yield less reliable scores. For example, the Bem Sex-Role Inventory (Bem, 1981), has higher reported score reliability on the short form as against the long form (Vacha-Haase, 1998). Regardless, unless the short form of a test contains items of equal or better quality than the long form, reliability will tend to decline.

Assume, for example, that a researcher has developed a 40-item data collection instrument that yielded scores (from her specific sample) with reliability of $\alpha = .80$. The researcher wishes to reduce the length of the test to 20 items, thereby reducing the time necessary to complete the instrument. We can use the Spearman-Brown prophecy formula (cf. Anastasi & Urbina, 1997) to estimate the impact of this reduction, which results in

marginal alpha of .67 (a .13 decrease). However, when the 40-item test yields scores with reliability that are perhaps acceptable but somewhat lower than the above example (say, α = .70) the impact on scores from the half-length short form is more meaningful. In this case alpha reduces to .54 (a .16 decrease), where only roughly one-half of the score variance is arguably true score variance (cf. Henson, 2000). This example also illustrates that the reduction in alpha is more dramatic when scores on the long form of the test yield lower reliability.

When developing short forms of tests, then, care should be taken to ensure the short form of the test contains enough items of sufficient quality to yield acceptable score reliability. However, this consideration must be viewed in light of the effect of sampling error on reliability estimates.

Exalting Sampling Error

As has been shown, the characteristics of the sample directly impact the reliability of the scores obtained. It can also be argued that other elements of measurement affect reliability estimates. Eason (1991, p. 84) indicated that "reliability is a characteristic of data," which suggests that all of the factors that can impact data (in this case scores on a test), can ultimately impact reliability. These factors may include but are not limited to: sample composition, testing

conditions, test-taker affect and fatigue, and/or time of measurement.

Whenever item deletion decisions are conducted to create short forms of tests, sample-based statistics (e.g., item difficulty and item discrimination coefficients) will almost always hinder the generalizability of the test's use with subsequent samples. A 25-item test reduced to 15 items may maintain appropriate score reliability for a particular sample, but it may not yield reliable scores with a different sample or even with the same sample under different conditions (test-retest reliability). Therefore, reliability should be estimated for scores from all samples. However, in absence of a reliability coefficient, Thompson and Vacha-Haase (2000) suggested:

> The crudest and barely acceptable minimal evidence of score quality in a substantive study would involve an explicit and direct comparison (Thompson, 1992) of (a) relevant sample characteristics (e.g., age, gender), whatever these may be in the context of a particular inquiry, with the same features reported in the manual for the normative sample or in earlier research and (b) the sample score SD with the SD reported in the manual or in other earlier research. (p. 190, emphasis in original)

Crocker and Algina (1986) agreed that "potential test users need to determine whether reliability estimates reported in test manuals are based on samples similar in composition and variability to the group for whom the test will be used" (p. 144, emphasis added).

Sampling error differences between samples are arguably less when these conditions are met. Psychometrically, coefficient alpha largely hinges on the total test variance, and therefore, will be more consistent between samples when the total test variances are comparable (holding all else constant in the alpha formula). Still, the best estimate of reliability for one's data is the actual reliability coefficient derived from one's data. With modern statistical software packages, this process takes at least a few seconds to complete.

An exception to the general expectation that reliabilities will vary between samples comes when a test developer reduces the number of items on a test using item statistics based on a normative sample and then re-administers the test to subsequent samples of sufficient size and whose characteristics parallel the original normative sample. Similarly, Dawis (1987) claimed that reliability "should be evaluated on a sample from the intended target population" (p. 486). In such cases, sampling error is arguably minimized and statistics are more likely to be stable (but not necessarily exactly the same) across studies.

However, this strategy is in stark contrast against how most short form tests are developed, particularly non-commercial data collection instruments often published in articles. While obvious logistical and cost limitations prohibit normative sampling in many, if not most cases, researchers should be aware of the risks to the integrity of scores obtained in a given administration of an instrument.

Of course, reliability estimates are not the only psychometric properties of scores dependent on sample characteristics. In efforts to parsimoniously reduce the number of items on a test, many researchers conduct exploratory and/or confirmatory factor analyses on the obtained scores (often from a non-normative sample) and delete items that do not behave as expected. Generally exploratory factor analyses are used to reduce a potential pool of items to smaller set of items for the factors of interest. Confirmatory factor analyses are often conducted to determine the theoretical stability of score structure with other samples and/or to establish construct validity for scores.

Like reliability, factor structure is a function of scores, is impacted by multiple factors, and is <u>not</u> solely the result of items on a test. Results of exploratory and/or confirmatory factor analyses may vary depending on sample characteristics and other conditions of measurement.

Sampling Error is as Sampling Error Does

Perhaps the following real-world example will help illustrate the potential impact of sampling error on both score reliability and factorial structure. The Teacher Efficacy Scale (TES; Gibson & Dembo, 1984) was used as a measure of general teaching efficacy (GTE) and personal teaching efficacy (PTE) in a study of preservice teachers (Henson, Stephens, & Grant, 1999). In its present form, the TES consists of 16 items that were reduced from 30 items by way of exploratory factor analysis in the Gibson and Dembo (1984) study. The TES has yielded generally reliable scores for both GTE (α = .64 to .77) and PTE (α = .75 to .81) in numerous prior studies (see e.g., Anderson, Greene,, Loewen, 1988; Hoy & Woolfolk, 1993; Moore & Esselman, 1992; Soodak & Podell, 1993). Historically, the TES has been used with such frequency that Ross (1994, p. 382) called it a "standard" instrument in the study of teacher efficacy. More recently, however, the utility of the TES has been questioned (cf. Henson, Bennett, Sienty, & Chambers, 2000; Tschannen-Moran, Woolfolk Hoy, & Hoy, 1998).

In the present example, the TES was given in pre and posttest format to 142 preservice teachers at a large state university in the southwest. Two weeks elapsed between administrations. With this design, it is possible to evaluate three important characteristics of the data: internal

consistency of scores from each scale (GTE and PTE) at both administrations via coefficient alpha, test-retest score reliability for both scales, and factor structure of scores for both administrations. Table 2 presents the reliability estimates from both administrations of the TES. It is important to remember that these statistics are based on responses to the same test from the same sample on two occasions separated by a two week delay.

INSERT TABLE 2 ABOUT HERE

Referring to Table 2, it appears that the PTE scale was more successful in yielding reliable scores at both pre and posttest. However, alphas varied for scores on both scales. Reliability decreased for GTE scores and increased for PTE scores between pre and posttest, with PTE showing the largest change. Interestingly, test-retest estimates indicated greater reliability for total GTE scores than PTE scores. These results highlight the impact of sample characteristics and context of measurement on obtained reliability estimates. There existed considerable fluctuation in reliability estimates by time of measurement. A comparison of alpha and test-retest estimates also revealed a reversal concerning which scale yielded the most reliable scores.

These results also suggest that in the classical test theory framework, these sources of measurement error are separate and cumulative (Anastasi & Urbina, 1997; Henson, 2000), a point too few researchers understand. That is, scores from a test may have error due to content sampling (internal consistency reliability) and separate error due to occasion of measurement (test-retest reliability) and separate error due to raters, if applicable (inter-rater reliability). Looking at the posttest and test-retest results for the present example, we can conceptualize the cumulative measurement error variance of the PTE scores to be [(1-.8133)+(1-.6621)=(.1867+.3379)=.5246. This leaves only about 48% of total score variance as true score variance, a result that is unacceptable by even the most psychometrically tolerant of researchers. As an aside, generalizability theory (as opposed to classical test theory) allows for the simultaneous examination of these sources of error as well as the interactions between them using ANOVA methodology. The interested reader is referred to Kieffer (1999) and Shavelson and Webb (1991) for accessible treatments of G theory.

Beyond score reliability, factor structure can also be impacted by sample characteristics or other elements of measurement. To illustrate this possibility, a principle components analysis with orthogonal rotation was conducted on both the pre and posttest administrations of the TES. Two

factors were extracted from both analyses based on the scree plot and prior research on the TES. Table 3 presents rotated factor pattern/structure coefficients for the two factors at both pre and posttest. For comparison purposes, the expected items for PTE factor are marked with an asterisk.

INSERT TABLE 3 ABOUT HERE

Looking at Table 3, the majority of items behaved as expected. However, items 13, 14, 15, and 16 were problematic (Note that this represents fully one quarter of all the items on the TES). Either these items did not weight at all (with a .30 threshold) on a factor, weighted on both factors (which is inconsistent with theoretical orthogonal solution), or weighted on the unexpected factor. Items 15 and 16 contained variance attributable to both factors at pretest as did item 14 at posttest. Item 16 failed to relate to either factor at posttest. Item 13 was particularly problematic in that it failed to weight on either factor at posttest and weighted on the unexpected factor at pretest.

Several points should be made concerning these results. First, a given test may not yield a factor structure identical to that obtained for a different sample. The present results illustrate the reality that factor structure inures to scores, not tests, and therefore is affected by sampling error. Second,

conditions of measurement may also impact score structure. Again, the present results varied from pre to posttest for the same sample. Any number of explanations are possible for these fluctuations (e.g., time of day, test-taker affect, etc). The point is clear, however, that score structure may vary by sample and even by administration. Scores are responses from people; people are different; people can change.

Implications for the Development of Short Form Tests

Regarding the development of short form tests, researchers should take care to evaluate their obtained score reliability and structure. Reliability analyses should be conducted in almost all cases (Thompson, 1994, 1999; Vacha-Haase, 1998). As noted by Pedhazer and Schmelkin (1991), "it is imperative to recognize that the relevant reliability estimate is the one obtained for the sample used in the [present] study under consideration" (p. 86, emphasis in original). Test developers should take caution when deleting items for at least two reasons as regards reliability. First, the loss of items may adversely impact subsequent reliability estimates since fewer items may negatively alter the obtained item to total variance ratio that is central to the alpha coefficient estimate. Second, researchers should be aware that the item statistics obtained are sample specific and may not (probably will not) hold for

future samples. Sampling error is a real threat to stable score reliability across studies.

Furthermore, score structure should also be examined whenever sample sizes permit some form of factor analysis. Items may (and probably will to some extent) behave differently with different subject pools, and as shown here, can do so even with the same subjects. Care should be taken when deleting items that do not behave as expected to reduce a test to a shorter, more parsimonious form. Ideally, multiple or normative samples should be considered when deleting borderline items.

Multiple factors should be considered when developing short form tests, whether they are commercially published tests or data collection instruments. Among these include: (a) the characteristics of the sample on which reliability estimates and score structure is based; (b) the assumed characteristics of the sample(s) intended for future test use; (c) and measurement conditions that may impact statistics, either in the "normative" sample or future samples. These sample characteristics and measurement conditions can be examined with reliability generalization methodology (Vacha-Haase, 1998). Reliability generalization, like validity generalization (Hunter & Schmidt, 1990; Schmidt & Hunter, 1977), can be used to characterize study features that impact reliability estimates across studies. The

reader is referred to Vacha-Haase (1998) for an introduction to reliability generalization.

These considerations are particularly salient for reducing test length based on data from non-normative samples, such as is often done with data collection instruments such as the TES described above. Very often items are deleted on the basis of results from a single administration of the test to one sample (often a sample of convenience). The TES, for example, was reduced from 30 items to 16 based on responses from 208 elementary school teachers in California (Gibson & Dembo, 1984). This test has subsequently been used with secondary, preservice, novice, special education, and expert teachers across the United States and even internationally (cf. Tschannen-Moran, Woolfolk Hoy, & Hoy, 1998). Needless to say, these groups may vary considerably from the original sample. Furthermore, several attempts have been made to shorten the TES even further. Hoy and Woolfolk (1993) developed a 10-item version that yielded acceptable score reliability (α = .72 for GTE; α = .77 for PTE) in their study. Although acceptable, these reliabilities are somewhat marginal and, as has been argued, may vary considerably with future samples. With only 10 items on the test, there is essentially less room for error. Hoy and Woolfolk rightfully recommended that other researchers always conduct factor analyses on their data but wrongfully assumed that the

responsibility for varying score structure between studies lies with the TES and not as a joint function of multiple factors, not the least of which is sampling error.

In sum, it is recommended that all studies estimate reliability for the scores at hand. This holds true for measurement and substantive studies since reliability inherently attenuates effect sizes (i.e., only systematic variance can be correlated between any two variables). The recently published report of the American Psychological Association Task Force on Statistical Inference (Wilkinson & APA Task Force on Statistical Inference, 1999) concurred, and recommended that authors "provide reliability coefficients of the scores for the data being analyzed even when the focus of their research is not psychometric" (p. 596).

Second, factor analyses should be conducted whenever possible to examine score structure (cf. Henson & Roberts, in press). Items should be deleted, and short form tests developed, only when the multiple factors that influence score reliability and structure are considered. Although shortened tests can save time, money, and effort, they may also come with a costly (albeit often overlooked) price of reduced score quality. Parsimony for the sake of parsimony is no replacement for informed researcher judgment concerning the psychometric integrity of scores.

References

Anastasi, A., & Urbina, S. (1997). Psychological testing (7th ed.) Upper Saddle River, NJ: Prentice Hall.

Anderson, R., Greene, M., & Loewen, P. (1988). Relationships among teachers' and students' thinking skills, sense of efficacy, and student achievement. Alberta Journal of Educational Research, 34, 148-165.

Bem, S. L. (1981). Bem Sex-Role Inventory: Professional manual. Palo Alto, CA: Consulting Psychologist Press.

Cohen, J. (1990). Things I have learned (so far). American Psychologist, 45, 1304-1312.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart and Winston.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16, 297-334.

Dawis, R. V. (1987). Scale construction. Journal of Counseling Psychology, 34, 481-489.

Eason, S. (1991). Why generalizability theory yields better results than classical test theory: A primer with concrete examples. In B. Thompson (Ed.), Advances in educational research: Substantive findings, methodological developments (Vol. 1, pp. 83-98). Greenwich, CT: JAI Press.

Gibson, S., & Dembo, M. H. (1984). Teacher efficacy: A construct

validation. Journal of Educational Psychology, 76, 569-582.

Henson, R. K., Stephens, J., & Grant, G. S. (1999, January). Self-efficacy in preservice teachers: Testing the limits of non-experiential feedback. Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio. (ERIC Document Reproduction Service No. ED 436 482)

Henson, R. K. (2000, November). A primer on coefficient alpha. Paper presented at the annual meeting of the Mid-South Educational Research Association, Bowling Green, KY.

Henson, R. K., Bennett, D. T., Sienty, S. F., & Chambers, S. M. (2000, April). The relationship between means-end task analysis and content specific and global self-efficacy in emergency certification teachers: Exploring a new model of self-efficacy. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Henson, R. K., & Roberts, J. K. (in press). Exploratory factor analysis reporting practices in published research. In B. Thompson (Ed.), Advances in social science methodology (Vol. 6). Stamford, CT: JAI Press.

Hoy, W. K., & Woolfolk, A. E. (1993). Teachers' sense of efficacy and the organizational health of schools. The Elementary School Journal, 93, 356-372.

Hunter, J. E., & Schmidt, F. L. (1990). Methods of meta-

analysis. Newbury Park, CA: Sage.

Kieffer, K. M. (1999). Why generalizability theory is essential and classical test theory is often inadequate. In B. Thompson (Ed.), Advances in social science methodology (Vol. 5, pp. 149-170). Stanford, CT: JAI Press.

Kuder, G. F., & Richardson, M. W. (1937). The theory of estimation of test reliability. Psychometrika, 2, 151-160.

Moore, W., & Esselman, M. (1992, April). Teacher efficacy, power, school climate and achievement: A desegregating district's experience. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Myers, D. G. (1990). Social psychology (3rd ed.). McGraw-Hill.

Pedhazer, E. J., & Schmelkin, L. P. (1991). Measurement, design, and analysis: An integrated approach. Hillsdale, NJ: Lawrence Erlbaum.

Reinhardt, B. (1996). Factors affecting coefficient alpha: A mini Monte Carlo study. In B. Thompson (Ed.), Advances in social science methodology (Vol. 4, pp. 3-20). Greenwich, CT: JAI Press.

Ross, J. A. (1994). The impact of an inservice to promote cooperative learning on the stability of teacher efficacy. Teaching and Teacher Education, 10, 381-394.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general

solution to the problem of validity generalization. Journal of Applied Psychology, 62, 529-540.

Shavelson, R., & Webb, N. (1991). Generalizability theory: A primer. Newbury Park, CA: Sage.

Soodak, L., & Podell, D. (1993). Teacher efficacy and student problem as factors in special education referral. Journal of Special Education, 27, 66-81.

Thompson, B. (1992). Two and one-half decades of leadership in measurement and evaluation. Journal of Counseling and Development, 70, 434-438.

Thompson, B. (1994). Guidelines for authors. Educational and Psychological Measurement, 54, 837-847.

Thompson, B. (1999, February). Understanding coefficient alpha, really. Paper presented at the annual meeting of the Education Research Exchange, College Station, TX.

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. Educational and Psychological Measurement, 60, 174-195.

Tschannen-Moran, M., Woolfolk Hoy, A., Hoy, W. K. (1998). Teacher efficacy: Its meaning and measure. Review of Educational Research, 68, 202-248.

Vacha-Haase, T. (1998). Reliability generalization: Exploring

variance in measurement error affecting score reliability across studies. Educational and Psychological Measurement, 58, 6-20.

Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. American Psychologist, 54, 594-604. (Reprint available through the APA Home Page: http://www.apa.org/journals/amp/amp548594.html)

Table 1

Hypothetical Data for Heterogeneous and Homogeneous Samples

Person/ Statistic	Item 1	2	3	4	5	Total Score
Heterogeneous Sample						
1	0	0	0	0	0	0
2	1	0	0	0	0	1
3	1	1	0	0	0	2
4	1	1	1	0	0	3
5	1	1	1	1	0	4
6	1	1	1	1	1	5
Item σ^2	.14	.22	.25	.22	.14	
Total σ^2						3.50
alpha						.83
Homogeneous Sample						
1	0	1	0	1	0	2
2	1	0	1	0	1	3
3	0	1	0	1	0	2
4	1	0	1	0	1	3
5	0	1	0	1	0	2
6	1	0	1	0	1	3
Item σ^2	.25	.25	.25	.25	.25	
Total σ^2						.30
alpha						-5.00

Note. This illustration is adapted from Reinhardt (1996) and Thompson (1999).

Table 2

Reliability Estimates for the Teacher Efficacy Scale (n = 142)

Variable	Pretest alpha	Posttest alpha	Test-Retest
GTE	.6456	.6102	.7094
PTE	.7423	.8133	.6621

Note. Test-retest reliability is based on total scores for both scales after two week delay. GTE = general teaching efficacy; PTE = personal teaching efficacy.

Table 3

Varimax Rotated Factor Pattern/Structure Coefficients for the Teacher Efficacy Scale at Pre and Posttest

Item No.	Pretest		Posttest	
	PTE	GTE	PTE	GTE
1*	.678		.664	
2		.549		.706
3		.489		.519
4		.553		.562
5*	.526		.661	
6*	.605		.734	
7*	.678		.709	
8		.777		.739
9*	.552		.701	
10*	.724		.709	
11		.418		.480
12*	.594		.596	
13*		.430		
14		.625	.480	.434
15*	.358	.337	.616	
16	.382	.311		

Note. * indicates items expected to associate with the PTE factor. Factor pattern/structure coefficients less than .30 are omitted.